Jurnal Barekeng
Vol. 6 No. 2 Hal. 45 – 50 (2012)

Barekeng
jurnal ilmu matematika dan terapan

PEMODELAN STATISTICAL DOWNSCALLING DENGAN PENDEKATAN REGRESI BAYES ROBUST PCA (STUDI KASUS : DATA GCM STASIUN AMBON)

FERRY KONDO LEMBANG
Staf Jurusan Matematika, FMIPA, Unpatti
Jl. Ir. M. Putuhena, Kampus Unpatti, Poka-Ambon
e-mail: ferrykondolembang@yahoo.co.id

ABSTRACT

Masalah mendasar dari prediksi prediksi model curah hujan adalah keakuratan model berdasarkan proses stokhastik skala global maupun skala kecil. *Statistical Downscalling* (SD) merupakan salah satu alternatif untuk mengatasi masalah tersebut. SD adalah model yang menghubungkan skala global GCM dengan skala yang lebih kecil (lokal) dengan jalan pra-pemrosesan .reduksi dimensi domain grid untuk mengatasi kasus multikolinearitas. Metode reduksi dimensi yang serikali digunakan adalah *Principal Component Analysis* (PCA). Namun PCA tidak dapat diandalkan jika ada pengamatan outlier dalam data, sehingga diperlukan reduksi dimensi yang *robust*. Reduksi dimensi *robust* menggunakan *Robust Principal Component Analysis* (ROBPCA) dengan estimator *robust* MCD. Dari hasil reduksi dimensi domain grid tersebut selanjutnya diregresikan dengan variabel respon berupa data curah hujan di stasiun Ambon dengan pendekatan regresi Bayes. Pendekatan regresi Bayes ROBUST PCA menjadi salah satu alternatif pada pemodelan SD. Hasil Penelitian menunjukkan Metode regresi Bayes ROBPCA cenderung lebih baik pada domain 8x8 dilihat pada kriteria kebaikan model RMSE terkecil yaitu 231,4 dan R-Square terbesar 38,1% dibandingkan domain 3x3 dan domain 12x12
Keywords: *Statistical Downscalling, GCM, ROBPCA, Regresi Bayes*

PENDAHULUAN

Analisis regresi merupakan analisis statistika yang bertujuan untuk memodelkan hubungan antara variabel bebas (X) dan variabel tidak bebas (Y). Metode *Ordinary Least Square* (OLS) merupakan salah satu metode estimasi parameter yang paling terkenal dalam model regresi karena relatih mudah. Kemudahan tersebut sebagai akibat adanya beberapa asumsi yang cukup ketat antara lain asumsi *error* identik independen dan berdistribusi normal yang harus dipenuhi sehingga akan diperoleh satu model taksiran untuk semua model data serta tidak terjadi kolinearitas ganda antara variabel bebas. Banyak metode estimasi parameter yang digunakan untuk mengatasi adanya multikolinearitas, antara lain: regresi komponen utama, regresi kuadrat terkecil parsial (PLS), regresi *ridge*, serta pendekatan regresi Bayes (Box and Tiao, 1973).

Salah satu penerapan yang dianggap sebagai penerapan pendekatan regresi Bayes dalam analisis regresi adalah Regresi *ridge*. Jika pada metode *Ordinary Least squares* (OLS) parameter regresi (β) diasumsikan konstan, tetapi pada pendekatan Bayes parameter model diasumsikan memiliki sebaran tertentu. Informasi ini disebut informasi prior. *Update* informasi prior pada parameter θ menggunakan informasi sampel yang terdapat dalam data (melalui fungsi *likelihood*), sehingga diperoleh informasi posterior yang akan digunakan untuk pengambilan keputusan (Gelman, dkk., 1995 dalam Prastyo, 2008). Prior pada regresi *ridge* adalah $\beta \sim N\left(\underline{\theta}, \sigma_\beta^2 I\right)$ yang berarti parameter regresi independen satu sama lain.

Pada beberapa kasus, korelasi diantara variabel independen terjadi dengan pola yang khusus (tertentu),

misalnya pada model curah hujan dengan data luaran GCM. Namun informasi GCM sifatnya global dan tidak berlaku untuk informasi skala kecil, sehingga untuk menjembatani Skala GCM ke Skala Kecil dipakai Teknik *Downscalling* (Wigena, 2006) yang merupakan teknik pereduksian dimensi. Metode reduksi dimensi dalam pra-pemrosesan yang digunakan antara lain : *Principal Component Análysis* (PCA) , Transformasi *Wavelet Diskrit* (TWD) (Anggraeni, 2009), Kernel PCA (Manorang, 2009), dan ROBUST PCA (Khotimah, 2009). Hasil reduksi dimensi dalam pra-pemrosesan menggunakan ROBUST PCA untuk mendapatkan validasi model curah hujan bisa diselesaikan dengan pendekatan regresi bayes sehingga dikenal dengan istilah regresi *Bayes ROBUST PCA*. Dalam perspektif statistika permasalahan ini merupakan pemodelan hubungan antara variabel iklim stasiun skala besar dengan komponen utama hasil reduksi pra-pemrosesan ROBUST PCA. Komponen utama hasil reduksidimensi ROBUST PCA dibagi atas 2 komponen yaitu, data *in-sample* untuk mendapatkan model dan data *out-sample* untuk mendapatkan validasi model. Kriteria kebaikan model untuk perbandingan kinerja hasil reduksi dimensi PCA dengan model regresi *Bayes ROBUST PCA* adalah RMSEP dan $R^2_{predict}$. Nilai RMSEP merupakan nilai dari *error* hasil taksiran sehingga model terbaik adalah model dengan RMSEP minimum yang menandakan nilai taksiran mendekati nilai sebenarnya sedangkan semakin besar nilai $R^2_{predict}$, maka semakin baik pula model yang didapatkan karena mampu menjelaskan lebih banyak data (Drapper dan Smith,1996).

TINJAUAN PUSTAKA

Principal Component Analysis (PCA)

PCA adalah suatu prosedur untuk mereduksi dimensi data dengan cara mentransformasi variabel-variabel asli yang berkorelasi menjadi satu set variabel baru yang independen yang merupakan kombinasi linier dari variabel asal sedemikian hingga varians menjadi maksimum (Johnson, 2002).

Misalkan vektor random $X^{'} = \left(X_1, X_2, ..., X_p\right)$ yang terdiri dari sejumlah observasi sebanyak p variabel dan mempunyai matriks varian-kovarian $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ mempunyai pasangan *eigenvalue-eigenvektor* $\left(\lambda_1 e_1\right), \left(\lambda_2 e_2\right), ..., \left(\lambda_p e_p\right)$, dimana $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

Maka kombinasi linier PC dapat ditulis sebagai berikut :

$$Z_1 = e_1 X = e_{11}X_1 + e_{21}X_2 + ... + e_{p1}X_p$$
$$Z_2 = e_2 X = e_{12}X_1 + e_{22}X_2 + ... + e_{p2}X_p$$
$$\vdots$$
$$Z_p = e_p X = e_{1p}X_1 + e_{2p}X_2 + ... + e_{pp}X_p$$

Model PC ke-i dapat juga ditulis dengan notasi $Z_i = e_i X$ dimana : i = 1,2,...,p dan oleh karenanya :

$$Var(Z_i) = e_i^{'} \Sigma e_i \qquad i = 1, 2, ..., p$$
$$Cov(Z_i, Z_k) = e_i^{'} \Sigma e_k \qquad i \neq k$$

PC tidak berkorelasi dan mempunyai varians yang sama dengan eigenvalue dari Σ, sehingga:

$$\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = tr(\Sigma) = \lambda_1 + \lambda_2 + ... + \lambda_p,$$

maka:

Proporsi varians ke-i = $\dfrac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p}$

Apabila PC yang diambil sebanyak k dimana (k<p), maka:

Proporsi variansi k PC = $\dfrac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$

Menurut Johnson (2002) dan Jollife (1986) ada beberapa acuan dalam menentukan banyaknya PC, yaitu: melihat scree plot, melihat *eigenvalue*yang lebih besar dari satu, dan total variansi yang dapat dijelaskan adalah 80 sampai 90 persen.

Pendeteksian *Outlier*

Outlier merupakan suatu pengamatan yang menyimpang cukup jauh dari pengamatan lainnya sehingga menimbulkan kecurigaan bahwa pengamatan tersebut berasal dari distribusi data yang berbeda (Hawkins dalam Sujatmiko, 2005:4). Pada data univariate, pengamatan *outlier* dapat dengan mudah terlihat dengan menggunakan beberapa plot sederhana, seperti scatter plot, steam and leaf, boxplot, dan sebagainya, sedangkan pada data multivariate identifikasi *outlier* umumnya didasarkan pada *Mahalanobis Distance* (MD),

$$d_{MD} = \sqrt{\left(x_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(x_i - \boldsymbol{\mu}\right)} \qquad (5)$$

dengan $\boldsymbol{\mu}$ merupakan vektor rata-rata data dan $\boldsymbol{\Sigma}$ merupakan matriks varian-kovarian. Suatu pengamatan diidentifikasi sebagai *outlier* jika suatu pengamatan mempunyai nilai d_{MD} lebih besar dari $\sqrt{\chi^2_{p,(1-\alpha)}}$. Namun identifikasi *outlier* pada data multivariate dengan jarak mahalanobis tidak maksimal karena adanya efek *masking* (adanya pengamatan *outlier* lain yang berdekatan) dan *swamping* (adanya pengamatan yang bukan outlier yang teridentifikasi sebagai *outlier*) (Rousseeuw dan Van Zomeren, 1990). Oleh karena itu, digunakan *Robust Distance* (RD) dengan estimator MCD(Rocke dan Woodruff, 1996), sehingga RD dapat dituliskan,

$$d_{RD} = \sqrt{\left(x_i - \mathbf{T(X)}_{MCD}\right)^T \mathbf{C(X)}^{-1}_{MCD} \left(x_i - \mathbf{T(X)}_{MCD}\right)} \qquad (6)$$

sama halnya dengan MD, sebuah pengamatan x_i diidentifikasi sebagai *outlier* jika mempunyai nilai d_{RD} lebih besar dari $\sqrt{\chi^2_{p,(1-\alpha)}}$.

Estimator MCD

Metode MCD merupakan upaya untuk menemukan h observasi ($h \leq n$) yang memiliki determinan matriks varian-kovarian terkecil dengan $[(n+p+1)/2] \leq h \leq n$.

$$MCD \approx \min\left\{\det\left(\mathbf{C(X)}_j\right)\right\}, \text{ j = 1, 2, ..., } \binom{n}{h}$$

di mana **C(X)** adalah matriks varian-kovarian berdasarkan pengamatan x_i dengan $i \in J$. Estimator MCD diberikan oleh: $\mathbf{T}(\mathbf{X}) = \frac{1}{h}\sum_{i=1}^{h} x_i$ dan

$$\mathbf{C}(\mathbf{X}) = \frac{1}{h-1}\sum_{i=1}^{h}\left(x_i - \mathbf{T}(\mathbf{X})\right)\left(x_i - \mathbf{T}(\mathbf{X})\right)^t$$

MCD mencari subsampel h, sebanyak ${}^nC_h$, sehingga untuk *n* besar dibutuhkan komputasi yang panjang untuk menemukan estimator MCD. Oleh karena itu, untuk meminimalisasi waktu komputasi digunakan algoritma FAST-MCD oleh Rousseeuw dan Van Driessen (1999). Inti dari algoritma FAST-MCD adalah *C-Step*.

Teorema*C-Steps*.

Diketahui $\mathbf{X} = \{\mathbf{x_1}, ..., \mathbf{x}_n\}^{\mathbf{t}}$ merupakan himpunan data sejumlah *n* observasi yang terdiri dari *p* variabel. Misal $H_1 \subset \{1, ..., n\}$ dimana $|H_1| = h.$ Tetapkan $\mathbf{T_1} := \left(\frac{1}{h}\right)\sum_{i\in H_1}^{h} \mathbf{x}_i$ dan $\mathbf{C_1} := \left(\frac{1}{h}\right)\sum_{i\in H_1}^{h}(\mathbf{x_i} - \mathbf{T_1})(\mathbf{x_i} - \mathbf{T_1})^t$. Jika det ($\mathbf{C_1}$)≠0 definisikan jarak relatif :

$$d_{1(i)} = \sqrt{(\mathbf{x_i} - \mathbf{T_1})^t\, \mathbf{C_1}^{-1}(\mathbf{x_i} - \mathbf{T_1})}\ ,\ i = 1, \ldots, n$$

Selanjutnya ambil himpunan H_2 sedemikian sehingga, $\{d_1(i); i \in H_2\} := \{(d_1)_{1:n}, ..., (d_1)_{h:n}\}$

di mana $(d_1)_{1:n} \le (d_1)_{2:n} \le \cdots \le (d_1)_{n:n}$ merupakan urutan jarak, kemudian $\mathbf{T_2}$dan $\mathbf{C_2}$dihitung berdasarkan himpunan H_2. Sehingga det $(\mathbf{C_2}) \le \det(\mathbf{C_1})$, akan sama jika dan hanya jika $\mathbf{T_1} = \mathbf{T_2}$ dan $\mathbf{C_1} = \mathbf{C_2}$. Tetapkan **T(X)** dan **C(X)** sebagai estimator dari subsampel yang memberikan determinan matriks varian-kovarian minimum. Berdasakan subsampel yang memberikan determinan matriks varian-kovarian minimum diberikan pembobotan pada data,

$$w_i = \begin{cases} 1 \text{ jika } (x_i - \mathbf{T(X)})^t\, \mathbf{C(X)}^{-1}(x_i - \mathbf{T(X)}) \le \chi^2_{p,0.975} \\ 0 \text{ lainnya} \end{cases}$$

Selanjutnya estimator MCD adalah:

$$\mathbf{T(X)}_{MCD} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \text{ dan}$$

$$\mathbf{C(X)}_{MCD} = \frac{\sum_{i-1}^{n} w_i (x_i - \mathbf{T(X)}_{MCD})(x_i - \mathbf{T(X)}_{MCD})^t}{\sum_{i=1}^{n} w_i - 1}$$

Regresi Linier

Analisis regresi adalah analisis statistika yang bertujuan untuk memodelkan hubungan antara variabel prediktan (respon) dengan variabel penjelas (Walpole, 1995).

Secara umum model yang menggambarkan hubungan antara variabel penjelas (X) dengan variabel respon (Y) adalah:

$$Y = f(X) + \varepsilon \qquad (12)$$

dalam bentuk matriks model regresi dinyatakan dengan (Draper dan Smith, 1992) :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Kriteria yang seringkali digunakan untuk kebaikan model regresi adalah RMSE dan R^2. Nilai RMSE menunjukkan keakuratan suatu model, sehingga model yang baik adalah model dengan nilai RMSE kecil. Nilai RMSE dari model dapat diperoleh dari persamaan:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n - p - 1}}$$

sedangkan R^2menunjukkan proporsi keragaman total nilai-nilai variabel respon yang dapat diterangkan oleh variabel-variabel prediktor dalam model yang digunakan. Secara umum, semakin besar nilai R^2 suatu model, maka model tersebut semakin baik. Nilai R^2 dapat dihitung dari,

$$R^2 = 1 - \frac{SS_{Error}}{SS_{\text{Total}}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Regresi Bayes

Model bayesian dikembangkan dari teorema bayes. Teorema bayes digunakan sebagai dasar dari metode penaksiran parameter suatu distribusi atau suatu model. Dalam teorema bayes, besaran parameter θ disajikan sebagai berikut :

$$p(\theta \mid \mathbf{x}) = \frac{L(\mathbf{x} \mid \theta)\, p(\theta)}{p(\mathbf{x})}$$

dengan $p(\theta)$ adalah distribusi prior, $L(\mathbf{x} \mid \theta)$ adalah likelihood dari sampel, dan $p(\theta \mid \mathbf{x})$ adalah distribusi posterior dari θ. Pembaharuan informasi prior pada parameter θ menggunakan informasi sampel yang terdapat dalam data (melalui fungsi likelihood), sehingga diperoleh informasi posterior yang akan digunakan untuk pengambilan keputusan.

Pendekatan Bayes dalam regresi dilakukan dengan membentuk sebaran posterior dari parameter (Lindley and Smith, 1972; Berger, 1985 dalam Setiawan, 2003). Posterior ini merupakan hasil kali antara prior dengan fungsi kemungkinan.

Model umum regresi normal ganda dengan k buah peubah bebas (termasuk intersep) adalah :

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

diasumsikan $\underline{\beta} \sim N(\underline{\theta}, V)$ dimana *V* adalah matriks ragam-peragam $\underline{\beta}$ sehingga simetris, sedangkan $\underline{y} \sim N\left(X\underline{\beta}, I\sigma^2\right)$. Dengan demikian fungsi priornya adalah :

$$p(\beta) \propto (2\pi)^{-k/2}\,|V|^{-1/2} \exp\left\{-\frac{1}{2}\left(\underline{\beta} - \underline{\theta}\right)^T V^{-1}\left(\underline{\beta} - \underline{\theta}\right)\right\}$$

Fungsi kemungkinan dari model regresi normal ganda adalah :

$$l(y|\beta) \propto \frac{1}{\sigma^n}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left(\underline{y}-X\underline{\beta}\right)^T\left(\underline{y}-X\underline{\beta}\right)\right] \quad (6)$$

GCM

GCM adalah suatu model berbasis komputer yang terdiri dari berbagai persamaan numerik dan deterministik yang terpadu dan mengikuti kaidah-kaidah fisika. Model ini menduga perubahan unsur-unsur cuaca dalam bentuk luaran grid-grid yang berukuran 100-500 km menurut lintang dan bujur (von Stroch *et al*. 1993 dalam Sutikno, 2008). GCM merupakan suatu alat penting dalam studi keragaman iklim dan perubahan iklim (Zorita dan Storch, 1999). Namun informasi GCM masih berskala global, sehingga sulit untuk memperoleh langsung informasi berskala lokal dari GCM. Tetapi GCM masih mungkin digunakan untuk memperoleh informasi skala lokal atau regional bila teknik *downscaling* digunakan (Fernandez, 2005 dalam Wigena, 2006).

Downscaling didefinisikan sebagai upaya menghubungkan antara sirkulasi variabel skala global (variabel penjelas) dan variabel skala lokal (variabel respon) (Sutikno, 2008). Untuk menjembatani skala GCM yang besar dengan skala yang lebih kecil (kawasan yang menjadi studi) digunakan teknik *Statistical Downscaling* (SD). SD adalah suatu proses *downscaling* yang bersifat statik dimana data pada grid-grid berskala besar dalam periode dan jangka waktu tertentu digunakan sebagai dasar untuk menentukan data pada grid berskala lebih kecil (Wigena, 2006).

Pendekatan SD menggunakan data regional atau global untuk memperoleh hubungan fungsional antara skala lokal dengan skala global GCM. Secara umum bentuk hubungan tersebut dinyatakan dengan:

$$\mathbf{Y = f(Z) + \varepsilon}$$

dimana:

Y : variabel respon (curah hujan)
Z :variabel penjelas (gabungan dari hasil reduksi spasial (lintang-bujur) variabel GCM)
ε : sisaan

METODOLOGI PENELITIAN

Data yang digunakan adalah data sekunder yang diperoleh dari data luaran GCM model CSIRO-Mk3 dari Australia, dengan domain GCM yang digunakan adalah domain 3x3 (9 grid), domain 8x8 (64 grid), dan domain 12x12 (144 grid). Lokasi grid yang diambil adalah ditengah-tengah Kabupaten Kota Ambon. Periode data yaitu tahun 1967-2000. Variabel yang digunakan pada penelitian ini adalah variabel luaran CSIRO Mk3 sebagai variabel prediktor yang meliputi: *precipitable water* (PRW), tekanan permukaan laut (SLP), komponen angin meridional (VA), komponen zonal (UA), ketinggian geopotensial (ZG), dan kelembaban spesifik (HUS). Ketinggian (level) yang digunakan dalam penelitian adalah 850 hPa, 500 hPa, dan 200 hPa. Sedangkan variabel respon yaitu data curah hujan bulanan Stasiun Kota Ambon. Adapun tahapan-tahapan analisis data dalam penelitian ini, yaitu :

1. Melakukan standarisasi data.
2. Mencari komponen utama menggunakan *principal component analysis* (PCA) dengan langkah seperti berikut:
   a. Membuat matriks varian-kovariansi **Σ**.
   b. Menurunkan nilai akar karakteristik (*eigen value*) λ dengan persamaan $|\mathbf{\Sigma} - \lambda \mathrm{I}| = 0$ dan *eigen vektor* dengan persamaan $\mathbf{\Sigma X} = \lambda_i \mathbf{X}$.
   c. Menetukan jumlah komponen utama yang dibangkitkan (dengan melihat keragaman kumulatif yang lebih besar sama dengan 85%).
   d. Mendapatkan variabel baru yaitu z_{CPCA}.

   Mencari komponen utama menggunakan *robust principal component analysis* (ROBPCA), dengan langkah seperti berikut:
   a. Menentukan elemen subsampel dari X, yaitu X_{h_1} yang diperoleh dari observasi terpilih.
   b. Menentukan $\mathbf{T}(\mathbf{X})_1$ dan $\mathbf{C}(\mathbf{X})_1$, $\det \mathbf{C}(\mathbf{X})_1$ dan $\mathrm{inv}\mathbf{C}(\mathbf{X})_1$.
   c. Menentukan RD_i.
   d. Mengurutkan nilai RD.
   e. Observasi yang mempunyai nilai RD terkecil ke-1 sampai dengan terkecil ke-h digunakan sebagai X_{h_2}
   f. Mengulang langkah b-d sampai diperoleh subsampel yang konvergen, yaitu $\det \mathbf{C}(\mathbf{X})_2 \leq \det \mathbf{C}(\mathbf{X})_1$. Tetapkan $\mathbf{T}(\mathbf{X})$ dan $\mathbf{C}(\mathbf{X})$ sebagai estimator subsampel yang mempunyai determinan matriks varian-kovarian minimum.
   g. Berdasarkan subsampel yang memberikan determinan matriks varian-kovarian minimum, diberikan pembobotan w_i terhadap data:
   h. Mendapatkan estimator MCD: $\mathbf{T}(\mathbf{X})_{MCD}$ dan $\mathbf{C}(\mathbf{X})_{MCD}$.
   i. Menentukan nilai akar karakteristik (*eigen value*) λ dengan menghitung $|\mathbf{C}(\mathbf{X})_{MCD} - \lambda \mathrm{I}| = 0$ dan *eigen vektor* dengan persamaan $\mathbf{C}(\mathbf{X})_{MCD}\mathbf{X} = \lambda_i \mathbf{X}$.
   e. Menentukan jumlah komponen utama yang dibangkitkan (dengan dengan melihat keragaman kumulatif yang lebih besar sama dengan 85%).
   f. Mendapatkan variabel baru yaitu z_{ROBPCA}.

3. Melakukan regresi linear berganda dengan variabel penjelas adalah komponen utama yang dihasilkan dari masing-masing metode, dengan model regresinya $Y = f(Z) + \varepsilon$.
4. Menganalisis kinerja hasil reduksi dimensi dan pemodelan SD dengan metode ROBUST PCA.

HASIL DAN PEMBAHASAN

Pra-pemrosesan Pemodelan SD

Langkah awal dalam pemodelan SD adalah reduksi dimensi, yang seringkali disebut sebagai pra-pemrosesan data. Pereduksian dilakukan pada dimensi spasialnya yaitu lintang dan bujur atau disebut grid dan pada semua variabel di setiap level serta pada setiap domain. Dalam hal ini setiap grid adalah variabel prediktor, sehingga pada domain 3x3, 8x8, dan 12x12secara berurutan ada 9, 64, dan 144 variabel yang akan direduksi.

Metode Reduksi Dimensi Robust PCA

Berdasarkan Kriteria beberapa komponen utama pertamanya telah menerangkan keragaman data lebih besar sama dengan 85% maka tabel 1 dibawah ini menerangkan PC optimal dari metode reduksi dimensi Robust PCA.

Tabel 1. Jumlah PC Optimal dan Keragaman Kumulatif PC Variabel Luaran GCM dengan Menggunakan Metode ROBPCA

No.	Variabel	Domain 3x3		Domain 8x8		Domain 12x12	
		Jml PC	Ker. Kum. (*)	Jml PC	Ker. Kum. (*)	Jml PC	Ker. Kum. (*)
1	HUSS	1	0.930	2	0.878	3	0.867
2	HUS200	1	0.986	1	0.853	2	0.902
3	HUS500	1	0.933	2	0.938	3	0.858
4	HUS850	1	0.920	2	0.887	3	0.867
5	PRW	1	0.937	2	0.889	2	0.899
6	SLP	1	0.920	1	0.876	1	0.961
7	UA200	1	0.955	1	0.912	2	0.973
8	UA500	1	0.923	2	0.865	3	0.900
9	UA850	1	0.934	2	0.856	2	0.855
10	VAS	1	0.917	2	0.897	3	0.852
11	VA200	1	0.924	1	0.943	2	0.889
12	VA500	1	0.856	3	0.857	5	0.860
13	VA850	1	0.923	3	0.898	4	0.853
14	ZG200	1	0.987	1	0.947	1	0.888
15	ZG500	1	0.995	1	0.966	1	0.896
16	ZG850	1	0.991	1	0.937	1	0.901

Berdasarkan Tabel 1 diketahui hasil reduksi dimensi variabel luaran GCM dengan menggunakan metode ROBPCA. Pada domain 3x3, jumlah komponen utama optimal yang terbentuk dengan keragaman yang dapat diterangkan lebih besar sama dengan 85% adalah satu komponen utama. Pada domain 8x8, komponen utama optimal yang digunakan antara satu sampai dengan tiga komponen utama. Pada domain 12x12, komponen utama optimal yang digunakan tidak lebih dari empat komponen utama, kecuali variabel HUSS dan VA500 yang menggunakan lima komponen utama.

Secara umum, variabel pada level permukaan mempunyai komponen utama yang semakin banyak sebanding dengan semakin luasnya domain, kecuali variabel SLP. Namun, hal tersebut juga tidak berlaku untuk variabel ZG200, ZG500, dan ZG850, karena cukup dengan satu komponen utama, variabel tersebut sudah mampu menjelaskan lebih dari 85% pada setiap domain. Berbeda dengan variabel HUSS, VAS, VA200, VA500, dan VA850 yang memerlukan cukup banyak komponen utama agar mampu menjelaskan lebih dari 85% total keragaman data.

Pemodelan SD

Tahap berikutnya adalah pemodelan SD. Pemodelan SD menggunakan regresi linier berganda, dengan variabel prediktor adalah gabungan dari variabel hasil reduksi dimensi variabel-varibel GCM pada masing-masing domain berdasarkan metode ROBUST PCA dan variabel respon yaitu data curah hujan bulanan stasiun Ambon.

Pemodelan SD dengan metode regresi Bayes ROBPCA menggunakan variabel prediktor yang merupakan gabungan dari variabel hasil reduksi dimensi variabel-variabel GCM dengan metode ROBPCA yang dilakukan pada setiap domain. Pada domain 3x3 menggunakan 16 variabel prediktor, pada domain 8x8 menggunakan 27 variabel prediktor, dan pada dom ain 12x12 menggunakan 38variabel prediktor (lihat Tabel 1). Nilai RMSE dan R^2 hasil pemodelan SD dengan menggunakan metode regresi bayes ROBPCA pada masing-masing stasiun dan domain teringkas dalam Tabel 2 berikut:

Tabel 2. RMSE dan R^2 Pemodelan SD dengan Metode Regresi BAYES ROBPCA

Stasiun Curah Hujan	GRID 3x3		GRID 8x8		GRID 12x12	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Ambon	245,9	28,2%	231,4	38,1%	244,43	36,4%

Berdasarkan Tabel 2 diketahui bahwa kinerja pemodelan SD antardomain tidak ada perbedaan yang signifikan untuk stasiun Ambon. Semakin luas domain semakin besar nilai R^2 dan semakin kecil nilai RMSE-nya. Nilai RMSE pada domain 8x8 ternyata lebih kecil dari nilai RMSE pada domain 3x3 dan 12x12. Hal ini berarti semakin luas domain tidak menjamin meningkatkan keakuratan suatu model dan sebaliknya. Hanya saja, untuk ukuran R-square terbesar 38,1 % ini belum dapat digolongkan model ini layak digunakan sebab kriteria layaknya model adalah $\geq 80\%$.

KESIMPULAN

Berdasarkan tujuan penelitian serta memperhatikan analisis dan pembahasan pada bab sebelumnya, maka diperoleh kesimpulan sebagai berikut:

1. Total variabel prediktor yang dihasilkan metode ROBPCA menurut domain secara berurutan adalah 16, 27, dan 38 variabel.
2. Pemodelan SD dilakukan dengan menggunakan regresi Bayes, dengan variabel prediktor adalah gabungan dari variabel hasil reduksi dimensi variabel GCM pada masing-masing domain berdasarkan metode ROBPCA dan variabel respon yaitu data curah hujan bulanan kota Ambon. Tidak terdapat konsistensi luasan domain terhadap besar kecilnya nilai RMSE dan R^2. Untuk kasus ini, model pada domain 8x8 menjadi yang paling baik sebab menghasilkan nilai RMSE terkecil dan R-Square terbesar.

DAFTAR PUSTAKA

Draper, N.R. dan Smith, H. (1992). *Analisis Regresi Terapan, Edisi kedua*. Jakarta: PT. Gramedia Pustaka Utama.

Johnson, R.A and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*. 5th Ed. New Jersey: Prentice Hall.

Jolliffe, I.T. (1986). *Principal Component Analysis, Second Ed*. New York: Springer-Verlag.

Rousseeuw, P.J. and Van Zomeren, B.C. (1990). "Unmasking Multivariate *Outlier*s and Leverage Points," *Journal of the American Statistical Association,* 85, 633–651.

Rousseeuw, P.J., and Van Driessen, K. (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, Vol. 41, No. 3, 212-223.

Sujatmiko, Irwan. (2005). "*Analisis Komponen Utama dengan Menggunakan Matriks Varians-Kovarians yang Robust"* Tesis. Jurusan Statistik-ITS. Surabaya.

Sutikno. (2008). "*Statistical Downscaling Luaran GCM dan Pemanfaatannya untuk Peramalan Produksi Padi*" Disertasi. Bogor: Program Pascasarjana, Institut Pertanian Bogor.

Walpole, R. E. (1995). " *Pengantar Statistika, Edisi ketiga*. Jakarta: PT. Gramedia Pustaka Utama.

Wigena, A.H. (2006). "*Pemodelan Statistical Downscaling dengan Regresi Projection Pursuit untuk Peramalan Curah Hujan Bulanan"* Disertasi. Bogor: Program Pascasarjana, Institut Pertanian Bogor.

Zorita, E. and von Storch, H., (1999): "The analog method as a simple statistical downscaling technique: comparison with more complicated method", *Journal of Climate*, 12, 2474-2489.